WPS4632

POLICY RESEARCH WORKING PAPER 4632

Instrumental Variables Regressions with Honestly Uncertain Exclusion Restrictions

Aart Kraay

The World Bank
Development Research Group
Macroeconomics and Growth Team
May 2008

Policy Research Working Paper 4632

Abstract

The validity of instrumental variable regression models depends crucially on fundamentally untestable exclusion restrictions. Typically exclusion restrictions are assumed to hold exactly in the relevant population, yet in many empirical applications there are reasonable prior grounds to doubt their literal truth. This paper shows how to incorporate prior uncertainty about the validity of the exclusion restriction into linear instrumental variable models, and explores the consequences for inference. In particular the paper provides a mapping from prior uncertainty about the exclusion restriction into increased uncertainty about parameters of interest. Moderate prior uncertainty about exclusion restrictions can lead to a substantial loss of precision in estimates of structural parameters. This loss of precision is relatively more important in situations where instrumental variable estimates appear to be more precise, for example in larger samples or with stronger instruments. These points are illustrated using several prominent recent empirical papers that use linear instrumental variable models.

This paper—a product of the Growth and the Macroeconomics Team, Development Research Group—is part of a larger effort in the department to develop tools for the analysis of development issues. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at akraay@worldbank.org.

Instrumental Variables Regressions with Honestly Uncertain Exclusion Restrictions

Aart Kraay
The World Bank

1818 H Street NW, Washington DC, 20433, akraay@worldbank.org. I would like to thank Daron Acemoglu, Laura Chioda, Frank Kleibergen, Dale Poirier and Luis Serven for helpful comments. The opinions expressed here are the author's and do not reflect the official views of the World Bank, its Executive Directors, or the countries they represent.

"The whole problem with the world is that fools and fanatics are always so certain of themselves, but wiser people are so full of doubts"

Bertrand Russell

The validity of the widely-used linear instrumental variable (IV) regression model depends crucially on the exclusion restriction that the error term in the structural equation of interest is orthogonal to the instrument. In virtually all applied empirical work this identifying assumption is imposed as if it held exactly in the relevant population. But in the vast majority of empirical studies using non-experimental data, it is hard to be certain that the exclusion restriction is literally true as it is fundamentally untestable.[1] Recognizing this, careful empirical papers devote considerable effort to selecting clever instruments and arguing for the plausibility of the relevant exclusion restrictions. But despite the best efforts of the authors, readers (and authors) of these papers may in many cases legitimately entertain doubts about the extent to which the exclusion restriction holds.

In this paper I consider the implications of replacing the standard identifying assumption that the exclusion restriction is literally true with a weaker one: that there is prior uncertainty over the correlation between the instrument and the error term, captured by a well-specified prior distribution centered on zero. The standard and stark prior assumption is that this distribution is degenerate with all of the probability mass concentrated at zero, so that the exclusion restriction holds with probability one in the population of interest. In most applications however a more honest, or at least more modest, prior assumption is that there is some possibility that the exclusion restriction fails, even if our best guess is that it is true.

I then explore the consequences for inferences about the structural parameters of interest of such prior uncertainty about the validity of the exclusion restriction. I find that even modest prior uncertainty about the validity of the exclusion restriction can lead to a substantial loss of precision in the IV estimator. Somewhat surprisingly, this loss of precision is relatively more important in situations in which the usual IV estimator would

[1] Murray (2006) poetically refers to this as the "cloud of uncertainty that hovers over instrumental variable estimation".

otherwise appear to be more precise, for example, when the sample size is large or the instrument is particularly strong. The intuition for this is straightforward. If I am willing to entertain doubts about the literal validity of the exclusion restriction, having a stronger instrument or having a larger sample size cannot reduce my uncertainty about the exclusion restriction, as the data are fundamentally uninformative about its validity. Since prior uncertainty about the exclusion restriction is unaffected by sample size or the strength of the instrument, while the variance of the IV estimator declines with sample size and the strength of the instrument for the usual reasons, the effects of prior uncertainty about the exclusion restriction become relatively more important in circumstances where the IV estimator would otherwise appear to be more precise.

In this paper I rely on the Bayesian approach to inference. With its explicit treatment of prior beliefs about parameters of interest, it provides a natural framework for considering prior uncertainty about the exclusion restriction. I use recently-developed techniques from the literature on Bayesian analysis of linear IV models, and extend them to allow for prior uncertainty over the validity of the exclusion restriction. However, to keep the results as familiar as possible (and hopefully as useful as possible) to non-Bayesian readers, I confine myself to particular cases that mimic standard frequentist results as closely as possible.

The broader goal of this paper is to provide a practical tool for producers and users of linear IV regression results who are willing to entertain doubts about the validity of their exclusion restrictions. Too often discussions of empirical papers that use IV regressions have an absolutist character to them. The author of the paper feels compelled to assert that the exclusion restriction relevant to his or her instrument and application is categorically true, and the skeptical reader, or seminar participant, or referee, is left in an uncomfortable "take it or leave it" position. One possibility is to wholeheartedly accept the author's untestable assertions regarding the literal truth of the exclusion restriction, and with them the results of the paper. The stark opposite possibility is to reject the literal truth of the exclusion restriction, and with it the results of the paper.

The results in this paper provide a modest but useful step away from such "foolish and fanatical" behaviour that the quote from Bertrand Russell reminds us of. For

example, using the results in this paper, the producers and consumers of a particular IV regression can readily agree on how much prior uncertainty about the validity of the exclusion restriction would be consistent with the author's results remaining significant at conventional levels. In some circumstances results might be quite robust to substantial prior uncertainty about the exclusion restriction, in which case the author and skeptical reader might agree that the author's conclusions are statistically significant even if they do not agree on the likelihood that the exclusion restriction is in fact true. In other circumstances, even a little bit of prior uncertainty about the exclusion restriction might be enough to overturn the significance of the author's results, in which case the reader who is skeptical about the validity of the exclusion restriction would be justified in rejecting the conclusions of the paper. The contribution of this paper is to provide an explicit tool to enable such robustness checks for uncertainty about the exclusion restriction.

I illustrate these results using three prominent studies that use linear IV regressions. Rajan and Zingales (1998) study the relationship between financial development and growth, using measures of legal origins and institutional quality as instruments for financial development. Frankel and Romer (1999) study the effects of trade on levels of development across countries, using the geographically-determined component of trade as an instrument. Finally, Acemoglu, Johnson and Robinson (2001) study the effects of institutional quality on development in a sample of former colonies, using historical settler mortality rates in the 18th and 19th centuries as instruments. In all three cases, reasonable readers might entertain some doubts as to the literal validity of the exclusion restriction. I show how to adjust the standard errors in core specifications from these papers to reflect varying degrees of uncertainty about the exclusion restriction. For the first two papers I find that moderate uncertainty about the exclusion restriction is sufficient to call into question whether the findings are indeed significant at conventional levels, while the findings of the third paper appear to be more robust to all but extreme prior uncertainty about the exclusion restriction.

Most theoretical and empirical work using the linear IV regression model proceeds from the assumption that the exclusion restriction holds exactly in the relevant population. One notable recent exception, closely related to this paper, is Hahn and Hausman (2006). They study the asymptotic properties of OLS and IV estimators when

there are known "small" violations of the exclusion restriction. In particular, they allow for a known correlation between the instrument and the error term, and in order to obtain asymptotic results they assume that this correlation shrinks with the square root of the sample size. Since the violation of the exclusion restriction is "local" in this particular sense, they find no effects on the asymptotic variance of the IV estimator. They then go on to compare the asymptotic mean squared error of the OLS and IV estimators, and show that IV dominates OLS according to this criteria unless violations of the exclusion restriction are strong. My approach and results differ importantly in two respects. First, I do not assume that the strength of violations of the exclusion restriction declines with sample size. While this assumption is analytically convenient when deriving asymptotic properties of estimators, it is not very intuitive. Since in general the data are uninformative about exclusion restrictions, it is unclear why we should think that concerns about the validity of the exclusion restriction are diminished in larger samples. Second, I explicitly incorporate uncertainty about the exclusion restriction, by assuming that there is a well-specified prior distribution over the correlation between the instrument and the error term. In contrast Hahn and Hausman (2006) treat violations of the exclusion restriction as a certain but unknown parameter to be chosen by the econometrician.[2] The uncertainty about the exclusion restriction that I emphasize is central to my results, as this uncertainty is responsible for the increased posterior uncertainty about parameters of interest. Closely related to their paper is Berkowitz, Caner and Fang (2008) who assume the same 'local' violation of the exclusion restriction, and demonstrate that standard test statistics in the IV regression model tend to over-reject the null hypothesis.

The results in this paper are also closely related to (although developed independently of) those in Conley, Hansen, and Rossi (2007). They study linear IV regression models in which there are potentially failures of the exclusion restriction (which they refer to as "plausible exogeneity"). They propose a number of strategies for investigating the robustness of inference in the presence of potentially invalid instruments, including a fully-Bayesian approach like the one taken here. While very similar in approach, this paper complements theirs in three respects. First, I focus on

[2] A similar approach of considering the sensitivity of coefficient estimates and tests of overidentifying restrictions to parametric violations of the exclusion restriction is taken by Small (2007).

special cases in which analytic or near-analytic results on the effects of prior uncertainty about the exclusion restriction are available, which helps to develop some key insights. In contrast, their paper uses numerical methods to construct and sample from the posterior distribution of the parameters of interest. Second, I characterize how the consequences for inference of prior uncertainty about the exclusion restriction depend on the characteristics of the observed sample. This can provide guidance to applied researches as to whether such prior uncertainty is likely to matter significantly in particular samples. Finally, I provide several macroeconomic cross-country applications of this approach that complement the more microeconomic examples in their paper.

The rest of the paper proceeds as follows. In order to develop intuitions based on analytic results, I begin in Section 2 with the simplest possible example of a bivariate OLS regression. This is of course a particular case of IV in which the regressor serves as its own instrument. I consider the consequences of introducing prior uncertainty about the correlation between the regressor and the error term for inference about the slope coefficient. In this simple case I can analytically characterize the effect of prior uncertainty of the precision of the OLS estimator. In Section 3 I turn to the IV regression model, focusing on the particular case of a just-identified specification with a single endogenous regressor. The same insights and analytic results from the OLS case apply to the OLS estimates of the reduced-form of the IV regression model. Although I am no longer able to analytically characterize the effect of prior uncertainty on the precision of the IV estimator of the structural slope coefficient of interest, it is straightforward to characterize it numerically and show how it depends on the characteristics of alternative realized samples. Section 4 of the paper applies these results to three empirical applications. Section 5 offers concluding remarks and discusses potential extensions of the results.

2. The Ordinary Least Squares Case

I begin by showing how to incorporate prior uncertainty about the exclusion restriction in the simplest possible case: a linear OLS regression. It is helpful to begin with this simple case by way of introduction. In the next section of the paper we will see how these results extend in a very straightforward way to linear IV regression models.

2.1 Basic Setup and the Likelihood Function

Consider the following bivariate linear regression:

(1) $$y_i = \beta \cdot x_i + \varepsilon_i$$

The regressor x is normalized to have zero mean and unit standard deviation. Assume further that the regressor and the error term are jointly normally distributed:

(2) $$\begin{pmatrix} \varepsilon_i \\ x_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma \cdot \rho \\ \sigma \cdot \rho & 1 \end{pmatrix} \right)$$

The key assumption here is that I allow for the possibility that the error term is correlated with the regressor, i.e. ρ might be different from zero. In the case of OLS this is the relevant failure of the exclusion restriction. In the next section when I discuss the IV case, I will assume that an instrumental variable z is available for x, but might be invalid in the sense that the instrument is correlated with the error term ε.

The distribution of the error term conditional on x is:

(3) $$\varepsilon_i | x_i \sim N\left(x_i \cdot \rho \cdot \sigma , \quad \sigma^2 \cdot \left(1 - \rho^2\right) \right)$$

Note of course that when $\rho \neq 0$, the usual conditional independence assumption $E[\varepsilon_i | x_i] = 0$ that is normally used to justify OLS does not hold.

Let y and X denote the Tx1 vectors of data on y and x in a sample of size T, and note that the normalization of x implies that X'X=T. Also let $\hat{\beta} = T^{-1}X'y$ denote the OLS estimator of the slope coefficient, and let $s^2 = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)/(T-1)$ be the OLS estimator of the variance of the error term. Finally, define $\omega \equiv \sigma^2 \cdot \left(1-\rho^2\right)$. With this notation the likelihood function can be written as:

$$(4) \quad L(y, X; \beta, \omega, \rho) \propto \omega^{-\frac{T}{2}} \cdot \exp\left[-\frac{1}{2} \cdot \left(\frac{(T-1)\cdot s^2}{\omega} + \frac{\left(\beta - \left(\hat{\beta} - \frac{\rho}{\sqrt{1-\rho^2}} \cdot \sqrt{\omega} \right) \right)^2}{\omega / T} \right) \right]$$

2.2 The Prior Distribution

In Bayesian analysis, the parameters of the model, in this case β, ω, and ρ, are treated as random variables. The analyst begins by specifying a prior probability distribution over these parameters, reflecting any prior information that might be available. This prior distribution for the parameters is then multiplied with the likelihood function, which is simply the distribution of the observed data conditional on the parameters. Using Bayes' Rule this delivers the posterior distribution of the model parameters conditional on the observed data sample. Inferences about the parameters of interest are based on this posterior distribution. In many applications, choosing an appropriately uninformative or diffuse prior distribution for the parameters results in a posterior distribution that is closely analogous to the usual frequentist results. In the case of a simple OLS regression where ρ=0 with certainty, an example of such a diffuse prior distribution is to assume that β and ln(ω) are independently and uniformly distributed, which implies that their joint prior distribution is proportional to 1/ω. In this case, a well-known textbook Bayesian result is that the marginal posterior distribution for β is a Student-t distribution with mean equal to the OLS slope estimate and variance equal to the estimated variance of the OLS slope. As a result, a standard frequentist 95

percent confidence interval would be analogous to the range from the 2.5th percentile to 97.5th percentile of the posterior distribution for β.

In order to retain this link with standard frequentist results, I will maintain this diffuse prior assumption for β and ω. My main interest is in specifying a non-degenerate prior distribution for the correlation between the regressor and the error term, ρ. Note that in the standard case there is a drastic asymmetry between prior beliefs about ρ and the other parameters of the model. In particular, prior beliefs about ρ are usually assumed to be highly informative in the sense that the prior probability distribution for ρ is degenerate with all the probability mass at zero, while prior beliefs about β and ω are assumed to be diffuse or totally uninformative. My objective is to relax this asymmetry by allowing for some prior uncertainty about the exclusion restriction. In particular, I assume the prior distribution for ρ is proportional to $\left(1-\rho^2\right)^{\eta}$ over the support (-1,1), where η is a parameter that governs prior confidence as to the validity of the identifying assumption. In particular, when η=0 we have a uniform prior over (-1,1). As η increases the prior becomes more concentrated around zero, and in the limit we approach the standard assumption that ρ=0 with probability one. Figure 1 plots this prior distribution for alternative values of η. The top panel of Table 1 reports the 5th and 95th percentiles of the distribution for alternative values of η. For example, setting η=500 corresponds to the rather strong prior belief that there is a 90 percent probability that ρ is between -0.05 and 0.05, and only a 10 percent probability that it is further away from zero.

A natural extension is to allow the prior distribution for ρ to have a non-zero mean, in order to encompass prior beliefs that there might be systematic violations of the exclusion restriction. Although this is straightforward to do, I do not pursue this option here as it adds little in the way of additional conceptual insights. For example, if our prior is that the mean of ρ is positive, then there will be a corresponding downward adjustment in the mean of posterior distribution for the slope coefficient. Moreover, the adjustments to the variance of the posterior distribution due to uncertainty about the exclusion restriction will be the same as what we have in the case where ρ has a zero mean, and these adjustments to the variance are of primary interest here.

Assuming further that the prior distribution for ρ is independent of the prior distribution for the other two parameters, we have the following joint prior distribution for the three parameters of the model:

(5) $$g(\beta,\omega,\rho) \propto \omega^{-1}\cdot\left(1-\rho^2\right)^{\eta}$$

2.3 The Posterior Distribution

The posterior density is proportional to the product of the likelihood and the prior density, i.e. from applying Bayes' Rule. Multiplying these two distributions and performing some standard rearrangements gives:

(6)
$$f\left(\beta,\omega,\rho \mid y,X\right) \propto \left(\omega/T\right)^{-\frac{1}{2}}\cdot\exp\left[-\frac{1}{2}\cdot\frac{\left(\beta-\left(\hat{\beta}-\frac{\rho}{\sqrt{1-\rho^2}}\cdot\sqrt{\omega}\right)\right)^2}{\omega/T}\right]$$
$$\cdot\ \omega^{-\frac{T+1}{2}}\cdot\exp\left[-\frac{1}{2}\cdot\left(\frac{(T-1)\cdot s^2}{\omega}\right)\right]\cdot\left(1-\rho^2\right)^{\eta}$$

The first line is proportional to a normal distribution for β conditional on ρ and ω, with mean $\hat{\beta}-\frac{\rho}{\sqrt{1-\rho^2}}\cdot\sqrt{\omega}$ and variance ω/T. When ρ=0, this is the very standard Bayesian result for the linear regression model with a diffuse prior. In particular, when ρ=0, the posterior conditional distribution of β is normal and is centered on the OLS estimate $\hat{\beta}$. When ρ is different from zero, the mean of the conditional posterior distribution for β needs to be adjusted to reflect this failure of the exclusion restriction. If the correlation between the regressor and the error term is positive (negative), then

intuitively, the posterior mean needs to be adjusted downwards (upwards) from the OLS slope estimator.

The second line is the joint posterior distribution of ω and ρ. It consists of the product of an inverted gamma distribution for ω and the posterior distribution for ρ.[3] The posterior distribution for ω is also standard, and intuitively has a mean equal to the OLS standard error estimator (times a small degress of freedom correction), i.e. $E[\omega]=\frac{(T-1)}{(T-3)}\cdot s^2$.

The only novel part of Equation (6) is the posterior distribution for ρ, which is identical to the prior distribution. This is what Poirier (1998) refers to as a situation in which the data are marginally uninformative about the unidentified parameter ρ. This in turn is a consequence of our prior assumption that ρ is independent of the other parameters of the model.[4] Although the data are uninformative about ρ, since we have now explicitly incorporated uncertainty about the exclusion restriction, we can explicitly average over this uncertainty when performing inference about the slope coefficient of interest, β. In particular, we know that the marginal posterior distribution of β will reflect our uncertainty about the exclusion restriction. We turn to this next.

2.3 Inference About β With an Uncertain Exclusion Restriction

Inferences about β are based on its marginal posterior distribution, which is obtained by integrating ρ and ω out of the joint posterior distribution of all three

---

[3] A random variable x follows an inverted gamma distribution, x~IG(α,β) if its pdf is: $f(x;\alpha,\beta)=\Gamma(\alpha)\cdot\beta^{-\alpha}\cdot x^{-(\alpha+1)}\cdot\exp\left(-\frac{1}{\beta\cdot x}\right)$. Setting $x=\sigma^2$, $\alpha=\frac{T-1}{2}$, $\beta=\frac{2}{s^2\cdot(T-1)}$ and disregarding the unimportant constant of proportionality gives the result in the text.

[4] If by contrast the prior distribution allowed for some dependence between the unidentified parameter ρ and the identified ones, then the posterior distribution for ρ would no longer be identical to the prior. Intuitively, if the unidentified and identified parameters are a priori dependent, then the data will through this channel be informative about the unidentified parameters.

parameters. This integration does not appear to be tractable analytically.[5] However, given the conditional structure of the posterior distribution, it is straightforward to compute the mean and variance of the marginal posterior distribution of β by repeated application of the law of iterated expectations. In particular, for the posterior mean we find:

$$(7) \quad E[\beta] = \hat{\beta} - s \cdot B(T) \cdot E\left[\frac{\rho}{\sqrt{1-\rho^2}}\right] = \hat{\beta}$$

where $B(T) \equiv \frac{\Gamma((T-2)/2)}{\Gamma((T-1)/2)} \cdot \sqrt{\frac{T-1}{2}} \to 1$ as T becomes large, and we have used the fact that $E\left[\sqrt{\omega}\right] = B(T) \cdot s$.

Note that the last expectation is with respect to the marginal posterior distribution of ρ. When ρ is identically equal to zero, we have the usual result that the mean of the posterior distribution of β is the OLS slope estimate. However, when there is prior (and thus also posterior) uncertainty about ρ, we have an additional term reflecting this uncertainty. This term involves the expectation (with respect to the posterior density for ρ) of $\frac{\rho}{\sqrt{1-\rho^2}}$. When the prior (and posterior) are symmetric around ρ=0, this term is unsurprisingly zero in expectation. If we are agnostic as to whether the correlation between the error term and x is positive or negative, on average this does not affect the posterior mean of β. Of course for other priors (and posteriors) not symmetric around zero this would not be the case, and the posterior mean of β would have to be adjusted accordingly.

The posterior unconditional variance is more interesting, and can also be found by repeated application of iterated expectations:

[5] When ρ=0, standard results show that integrating ω out of the joint posterior distribution results in a marginal t-distribution for β. However this convenient standard result does not go through when ρ differs from zero.

(8) $$V[\beta] = s^2 \cdot \left(\frac{1}{T} + E\left[\frac{\rho^2}{1-\rho^2} \right] \right) \cdot \left(\frac{T-1}{T-3} \right)$$

Disregarding the small degrees of freedom correction $(T-1)/(T-3)$, the first term is just the standard OLS estimator of the variance of $\hat{\beta}$, which is $\frac{s^2}{T}$. The second term is a correction to the variance estimator coming from the fact that there is uncertainty about the conditional mean of β coming from our uncertainty about ρ. In fact, the second term is recognizeable as the variance of the adjustment to the conditional mean that we saw above.

This correction to the posterior variance of β is quantitatively very important because it does not decline with the sample size T. The reason for this is straightforward -- since the data are uninformative about the correlation between the regressor and the error term, having a larger sample cannot reduce our uncertainty about this parameter.

The bottom panel of Table 1 gives a sense of the quantitative importance of this adjustment to the posterior variance. Define $\left(1 + T \cdot E\left[\frac{\rho^2}{1-\rho^2} \right] \right)^{1/2}$ as the ratio of the standard deviation of the posterior distribution of β in the case where there is prior uncertainty about ρ, to the same standard deviation in the standard case where ρ is identically equal to zero. This ratio captures the inflation of the posterior standard deviation due to uncertainty about ρ. This ratio can be large, particularly in cases where the sample size is large and/or when there is greater prior uncertainty about ρ. For example, for the case where η=100, so that 90 percent of the prior probability mass for ρ lies between -0.12 and 0.12, the posterior standard deviation is 22 percent higher in a sample size of 100, but 87 percent higher when the sample size is 500, and 245 percent larger in a sample of size 1000. Moving to the left in the table to cases with greater prior uncertainty about ρ results in even greater inflation of the posterior standard deviation.

In summary, in this section I have shown how to incorporate prior uncertainty about the relevant exclusion restriction in a very simple OLS example. The main insight from this section is that even modest doses of prior uncertainty about the exclusion restriction can substantially magnify the variance of the posterior distribution of β. Moreover, this effect is greater the larger is the sample size, as the intrinsic uncertainty about the exclusion restriction becomes relatively more important. The results of this section will be helpful in developing results for the IV case in the following section, and the key insight regarding the role of sample size will generalize naturally.

3 The Instrumental Variables Case

I now extend the results of the previous section to the case of the linear IV regression model in which there is prior uncertainty about the validity of the exclusion restriction. In this section I show that this type of uncertainty magnifies the posterior variance of the slope coefficients in the reduced-form version of the model, and this in turn makes the unconditional posterior distribution of the structural slope coefficient of interest more dispersed. I also show how this increase in dispersion depends on the characteristics of the observed sample.

3.1 Basic Setup

To keep things as simple as possible I focus on the particular case where the dependent variable y is a linear function of a single potentially endogenous regressor, x, and a single instrument z is available for x. The structural form of the model is:

$$\begin{aligned} y_i &= \beta \cdot x_i + \varepsilon_i \\ x_i &= \Gamma \cdot z_i + v_i \end{aligned} \tag{9}$$

The main parameter of interest is β, which captures the structural relationship between y and x. The parameter Γ captures the relationship between the instrument z and the endogenous variable x.

For convenience I assume that, like the endogenous regressor x, the instrument z has also been normalized to have a zero mean and unit standard deviation. I assume further that the two error terms and the instrument are jointly normally distributed:

$$\begin{pmatrix} \varepsilon_i \\ v_i \\ z_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \sigma_\varepsilon^2 & \lambda \cdot \sigma_\varepsilon \cdot \sigma_v & \rho \cdot \sigma_\varepsilon \\ & \sigma_v^2 & 0 \\ & & 1 \end{pmatrix} \right) \tag{10}$$

where σ_ε^2 and σ_v^2 are the variances of the two error terms, and λ and ρ are the correlations of ε with v, and ε with z, respectively.

The standard assumption used to identify the linear IV model is that the correlation ρ between the instrument z and the error term ε is identically equal to zero. This is the exclusion restriction which stipulates that the only channel through which the instrument z affects the dependent variable y is through the endogenous variable x. When the exclusion restriction holds, it is possible to separate the the regressor x into (i) an endogenous component, v, that has a potentially nonzero correlation λ with the error term, and (ii) an exogenous component $\Gamma \cdot z$ that is uncorrelated with the error term when $\rho=0$. This latter exogenous source of variation in x can then be used to identify the slope coefficient β. In fact, this is precisely the intuition behind two-stage least squares (2SLS) estimation. In the first stage, the endogenous variable is regressed on the instrument x. The fitted values from this first-stage regression are used as a proxy for the exogenous component of x in the second-stage regression.

When the exclusion restriction fails to hold, the instrumental variables estimator of β is biased with a bias equal to $\frac{\rho \cdot \sigma_\varepsilon}{\Gamma}$. This bias is larger (in absolute value) the larger is the correlation between the instrument and the error term, and the weaker is the correlation between the instrument and the endogenous variable x, i.e. the smaller is Γ.

Standard practice is to impose the identifying assumption and proceed as if it were literally true. This approach is appealing because it ensures -- albeit purely by assumption -- that the IV estimator will be consistent for β. But in most empirical applications using non-experimental data, it is impossible to be sure that the exclusion restriction in fact holds, as it is fundamentally untestable.

Bayesian analysis of the linear IV model is most conveniently based on the reduced form of the model in Equation (9). The reduced form is obtained by substituting the second equation into the first:

$$\begin{aligned} y_i &= \gamma \cdot z_i + u_i \\ x_i &= \Gamma \cdot z_i + v_i \end{aligned} \tag{11}$$

where $u_i \equiv \varepsilon_i + \beta \cdot v_i$ and $\gamma \equiv \beta \cdot \Gamma$. This latter identity allows us to retrieve the slope parameter of interest, β, from the coefficients of the reduced-form model. This is precisely the principle of indirect least squares. In particular, in the just-identified case I consider here, the 2SLS estimator of β is the ratio of the OLS estimators of γ and Γ from the two equations of the reduced form.

The distributional assumptions for the structural form of the model imply the following distribution for the reduced form errors and the instrument:

$$(12) \qquad \begin{pmatrix} u_i \\ v_i \\ z_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \sigma_u^2 & \theta \cdot \sigma_u \cdot \sigma_v & \phi \cdot \sigma_u \\ & \sigma_v^2 & 0 \\ & & 1 \end{pmatrix} \right)$$

where:

$$(13) \qquad \begin{aligned} \sigma_u^2 &= \sigma_\varepsilon^2 + 2\beta\lambda\sigma_\varepsilon\sigma_v + \beta^2\sigma_v^2 \\ \theta &= \frac{\lambda\sigma_\varepsilon + \beta\sigma_v}{\sqrt{\sigma_\varepsilon^2 + 2\beta\lambda\sigma_\varepsilon\sigma_v + \beta^2\sigma_v^2}} \\ \phi &= \frac{\rho\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + 2\beta\lambda\sigma_\varepsilon\sigma_v + \beta^2\sigma_v^2}} \end{aligned}$$

Note that the correlation ϕ between the reduced form error u and the instrument z is the counterpart of the correlation ρ between the structural form error ε and the instrument z. When the exclusion restriction holds exactly, $\rho=\phi=0$, and we have the standard linear IV regression model. In the next section of the paper I show how to replace this exact exclusion restriction with something weaker: a non-degenerate prior probability distribution over the correlation between the instrument and the error term.

The distribution of the reduced-form errors u and v conditional on the instrument is:

$$(14) \qquad \begin{pmatrix} u_i \\ v_i \end{pmatrix} \Big| z_i \sim N\left(\begin{pmatrix} \phi \cdot \sigma_u \cdot z_i \\ 0 \end{pmatrix} , \begin{pmatrix} \sigma_u^2 \cdot (1-\phi^2) & \theta \cdot \sigma_u \cdot \sigma_v \\ & \sigma_v^2 \end{pmatrix} \right)$$

This in turn implies the following distribution for y and x conditional on the instrument:

$$(15) \qquad \begin{pmatrix} y_i \\ x_i \end{pmatrix} \Big| z_i \sim N\left(\begin{pmatrix} (\gamma + \phi \cdot \sigma_u) \cdot z_i \\ \Gamma \cdot z_i \end{pmatrix} , \begin{pmatrix} \sigma_u^2 \cdot (1-\phi^2) & \theta \cdot \sigma_u \cdot \sigma_v \\ & \sigma_v^2 \end{pmatrix} \right)$$

Let Y denote the Tx2 matrix with the T observations on (y_i,x_i) as rows; let Z denote the Tx1 vector containing the T observations on z_i; and recall that Z has been normalized such that Z'Z=T. Let Ω denote the variance-covariance matrix of (y_i,x_i) conditional on z_i. Define the 1x2 matrix $G \equiv (\gamma : \Gamma)$ and let $\hat{G} \equiv (\hat{\gamma} : \hat{\Gamma}) = (Z'Z)^{-1}Z'Y$ denote the matrix of OLS estimates of the reduced-form slope coefficients and $S \equiv (Y - Z\hat{G})'(Y - Z\hat{G})/(T-1)$ as the estimated variance-covariance matrix of the residuals from the OLS estimation of the reduced-form slopes. The multivariate generalization of the likelihood function in Equation (4) is:[6]

$$(16) \qquad \begin{aligned} L(Y,X,Z;G,\Omega,\phi) &= \pi^{-TM/2} \cdot |\Omega|^{-T/2} \\ &\cdot \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \Omega^{-1} \left((T-1) \cdot S + T \cdot \left(G - \left(\hat{G} - (\phi \cdot \sigma_u : 0)\right)\right)' \left(G - \left(\hat{G} - (\phi \cdot \sigma_u : 0)\right)\right) \right) \right\} \right] \end{aligned}$$

3.2 Bayesian Analysis of the IV Regression Model

When the exclusion restriction holds exactly, i.e. $\rho=\phi=0$, the reduced-form model in Equation (11) becomes a standard multivariate linear regression model, in this particular case with two equations in which the dependent variable y and the endogenous regressor x are both regressed on the instrument z. Bayesian analysis of

[6] See for example Zellner (1973), Equation 8.6 or Poirier (1996), Equation 10.3.12.

the linear IV model builds on well-established textbook results for Bayesian analysis of the multivariate regression model (for textbook treatments of the latter see Zellner (1971), Ch. 8 and Poirier (1996), Ch. 10). In particular, the multivariate regression model admits a natural conjugate prior, meaning that the prior and posterior distributions have the same analytic form. Moreover, there are analytic results providing the mapping from the parameters of the prior distribution to the parameters of the posterior distribution, which make transparent how the observed data is used to update prior beliefs.

Hoogerheide, Kleibergen and Van Dijk (2008) extend these tools to analysis of the linear IV regression model. Their key insight is that, since there is a one-to-one mapping between the structural and the reduced-form parameters, the familiar prior and posterior distributions for the reduced form parameters in the multivariate regression model induce well-behaved prior and posterior distributions over the structural parameters. They analytically characterize these distributions for the structural parameters for a number of particular cases, and provide an application to the Angrist-Krueger data. I follow their approach, but with a further extension to allow for prior uncertainty over the validity of the exclusion restriction.

3.3 The Prior Distribution

I begin by specifying the same prior distribution over the correlation between the reduced-form error and the instrument, ϕ, that was used in the previous section of the paper, i.e. $g(\phi) \propto (1-\phi^2)^{\eta}$ over the support (-1,1) where η is a parameter that governs the strength of the prior belief that this correlation is zero. For the remaining parameters, I make the standard multivariate analog of the diffuse prior assumptions for these parameters in the OLS case. In particular, define $\omega_{11} = (1-\phi^2)\cdot\sigma_u^2$, $\omega_{12} = \theta\cdot(1-\phi^2)^{-1/2}$, and $\omega_{22} = \sigma_v^2$, so that $\Omega = \begin{pmatrix} \omega_{11} & \omega_{12}\cdot\omega_{11}\cdot\omega_{22} \\ & \omega_{22} \end{pmatrix}$, and let the prior distribution for the elements of Ω be $|\Omega|^{-3/2}$. This prior corresponds to the Jeffrey's prior for the multivariate regression model when ϕ=0. And, as in the OLS case, this choice of prior distribution ensures that the Bayesian results mimic the frequentist ones for the

case where ϕ=0. In this case, the posterior distribution for the reduced-form slopes is a multivariate Student-t distribution centered in the OLS slope estimates. With the further assumption that the prior distribution of the reduced-form slopes is uniform and independent of the other parameters, we have the following joint prior distribution:

$$(17) \qquad g(G,\Omega,\phi) \propto |\Omega|^{-3/2} \cdot (1-\phi^2)^{\eta}$$

Before proceeding, it is useful to characterize the prior distribution that this implies for the correlation between the structural disturbance and the instrument, ρ.

Since $\rho = \dfrac{\phi\,\sigma_u}{\sqrt{\sigma_u^2 - 2\,\theta\,\sigma_u\sigma_v\gamma/\Gamma + (\gamma/\Gamma)^2\sigma_v^2}}$, the prior distribution of ρ will in general depend on the entire joint prior distribution of all of the structural parameters. However, since the prior distribution of the remaining parameters is chosen to be uninformative, it is straightfoward to verify numerically that the distribution of ρ has the same shape and percentiles as the distribution of ϕ.[7] As a result, we can use the percentiles reported in Table 1 for the prior distribution of ρ in the OLS case to interpret the prior distributions of ϕ and ρ in the IV case.

3.4 The Posterior Distribution

The posterior distribution for the parameters of interest is proportional to the product of the likelihood function and the prior, i.e. from applying Bayes' Rule. Multiplying these two distributions and rearranging gives:

[7] It is straightforward although tedious to compute the Jacobian of the mapping from the structural parameters to the reduced-form parameters, and use this to write down the joint prior distribution of all the structural-form parameters. It does not however appear to be tractable to extract analytically from this the implied marginal distribution of ρ. This is why I instead characterize this distribution numerically.

$$L(G,\Omega,\phi|Y,X,Z)\propto$$

$$(18)\qquad \cdot\left|\Omega\right|^{-1/2}\cdot\exp\left[-\frac{1}{2}\left(G-\left(\hat{G}-\left(\frac{\phi\cdot\omega_{11}}{\sqrt{1-\phi^2}}:0\right)\right)\right)'\left(\frac{\Omega}{T}\right)^{-1}\left(G-\left(\hat{G}-\left(\frac{\phi\cdot\omega_{11}}{\sqrt{1-\phi^2}}:0\right)\right)\right)\right]$$

$$\cdot\left|\Omega\right|^{-(T-2)/2}\cdot\exp\left[-\frac{1}{2}\mathrm{tr}\left\{\Omega^{-1}S\cdot(T-1)\right\}\right]\cdot\left(1-\phi^2\right)^{\eta}$$

This expression is just the multivariate generalization of Equation (6). The first line is proportional to a normal distribution for the matrix of reduced-form slopes, G, conditional on ϕ and Ω, with mean $\left(\hat{\gamma}-\frac{\phi\cdot\omega_{11}}{\sqrt{1-\phi^2}}\ :\ \hat{\Gamma}\right)$ and variance-covariance matrix $\frac{\Omega}{T}$. When ϕ=0, we again retrieve the standard Bayesian result for the multivariate linear regression model with a diffuse prior for the reduced-form of the IV regression. In particular, when ϕ=0, the posterior conditional distribution of the reduced-form slopes is normal and is centered on their OLS estimates. However, when ϕ is different from zero, the mean of the conditional posterior distribution for γ needs to be adjusted to reflect this failure of the exclusion restriction, which induces a correlation between the regressor and the error term in the first structural equation. If the correlation between the regressor and the error term is positive (negative), then intuitively, the posterior mean needs to be adjusted downwards (upwards) from the OLS slope estimator. In contrast, no adjustment is required for the conditional mean of Γ, since by assumption the error term in the second structural equation is orthogonal to the instrument.

The second line is the joint posterior distribution of Ω and ϕ, and is again precisely analogous to the OLS case. It consists of the product of an inverted Wishart distribution for Ω and the posterior distribution for ϕ. The posterior inverted Wishart distribution for Ω is the multivariate generalization of the inverted gamma distribution for ω in the OLS case, and again it is intuitively centered on the OLS variance estimator, i.e. $E[\Omega]=\frac{T-1}{T-3}\cdot S$.

As in the OLS case, the only novel part of Equation (6) is the posterior distribution for ϕ, which once again is identical to the prior distribution. As before, the prior and the posterior are identical because the data are marginally uninformative about this parameter given the prior independence between ϕ and the other parameters of the model. However, since we have explicitly incorporated uncertainty about the exclusion restriction, we can explicitly average over this uncertainty when performing inference about the slope coefficients of interest.

3.4 Inference with an Uncertain Exclusion Restriction

As in the OLS case, we want to base inferences about β on its marginal posterior distribution, which is obtained by integrating all of the other parameters out of the joint posterior distribution. Again, this is unfortunately not tractable analytically and needs to be done numerically. However, we can obtain some useful insights by first studying how the distribution of the reduced-form slopes is affected by prior uncertainty about the exclusion restriction.

We begin by using the law of iterated expectations to compute the unconditional posterior mean and variance of the reduced-form slopes:

$$(19) \quad E[\gamma : \Gamma] = \left(\hat{\gamma} - s \cdot B(T) \cdot E\left[\frac{\phi}{\sqrt{1-\phi^2}} \right] : \hat{\Gamma} \right) = \left(\hat{\gamma} : \hat{\Gamma} \right)$$

and

$$(20) \quad V[\gamma : \Gamma] = \begin{pmatrix} s_{11}^2 \left(1/T + E\left[\phi^2/(1-\phi^2)\right]\right) & s_{12}/T \\ s_{12}/T & s_{22}^2/T \end{pmatrix} \cdot \left(\frac{T-1}{T-3} \right)$$

These expressions are just the multivariate generalizations of Equations (6) and (7) in the OLS case, and the intuitions for them are identical. Since the prior (and posterior) distribution for ϕ has zero mean, the expectation in Equation (19) is equal to zero and so the unconditional posterior mean for the reduced-form slopes is equal to their OLS

estimates. The effects on the posterior variance are substantively more interesting. As before, we see that the posterior variance of γ increases due to uncertainty about the exclusion restriction. In fact, the posterior variance of γ is identical to the OLS case. It consists of the usual component that declines with sample size, s_{11}^2 / T, as well as an adjustment capturing the variance of the adjustment to the sample mean due to uncertainty about the exclusion restriction, $s_{11}^2 \cdot E\left[\phi^2 / (1-\phi^2)\right]$. The key point once again is that this adjustment does not decline with sample size, and so uncertainty about the exclusion restriction has proportionately larger effects on the posterior variance of the reduced-form slope coefficient γ when the sample size is large. In contrast, there is no change in the posterior variance of the slope coefficient from the first-stage regression, Γ, as the exclusion restriction is not relevant to the estimation of this slope parameter.

This adjustment to the posterior variance of the reduced-form coefficient γ will also be reflected in the distribution of the structural form coefficients. In particular, since $\beta=\gamma/\Gamma$, and since uncertainty about the exclusion restriction expands the posterior variance of γ alone, we would expect to see a similar increase in the dispersion of the posterior distribution of β as well. I characterize this effect by sampling from the posterior distribution of β. In fact, since the posterior distribution of β conditional on ϕ and Ω is a Cauchy-like ratio of correlated normal random variables, it is not even clear that moments of the unconditional posterior distribution of β exist.

In general, the effects of prior uncertainty about the exclusion restriction on the posterior distribution of the structural slope coefficient of interest will be sample-dependent. This is because the posterior distribution in Equation (18) depends on the observed sample through the OLS estimates of the reduced form slopes and residual variances, $\hat{\gamma}$, $\hat{\Gamma}$, and S. In order to give a sense of how the effects of prior uncertainty about the exclusion restriction might vary in different observed samples, I present some simple illustrative calculations for alternative hypothetical observed samples. I begin by innocuously assuming that the observed data on y and x are scaled to have mean zero and variance one, as is z. The observed sample can therefore be characterized by three sample correlations, R_{yx}, R_{yz}, and R_{xz}, and the observed reduced-form slopes and residual variances can be expressed in terms of these correlations as:

(21) $\left(\hat{\gamma}:\hat{\Gamma}\right)=\left(R_{yz}:R_{xz}\right)$ and $S=\begin{pmatrix} 1-R_{yz}^2 & R_{xy}-R_{zy}\cdot R_{xz} \\ & 1-R_{xz}^2 \end{pmatrix}$

For each hypothetical sample summarized by a combination of assumptions on the three sample correlations, I sample from the posterior distribution of β, for a range of values for the parameter governing prior uncertainty about the exclusion restriction, η, and for different values of the sample size, T. I take 10,000 draws from the posterior distribution of β in each case, and compute the 2.5th and 97.5th percentiles of the distribution. This is analogous to a standard frequentist 95 percent confidence interval for the IV estimate of the slope coefficient.

The results of this exercise are summarized in Table 2. Each row of the table corresponds to a set of assumptions on the observed sample correlations and the sample size. These assumptions are spelled out in the left-most columns, in italics. In each row I also report the 2.5th and 97.5th percentiles of the posterior distribution for β in the standard case where there is no uncertainty about the exclusion restriction, i.e. when $\rho=\phi=0$. This serves as a benchmark. The right-most columns correspond to various assumption about η, corresponding to varying degrees of prior certainty about the exclusion restriction. I consider the same range of values as in Table 1, and for reference at the top of the table I report the 5th and 95th percentiles of the prior distribution of ϕ (and ρ) that these imply. Each cell entry reports the length of the interval from the 2.5th to the 97.5th percentile of the posterior distribution of β, expressed as a ratio to the length of this same interval when $\rho=\phi=0$, i.e. relative to the standard case.

Not surprisingly, all of the entries in Table 2 are greater than one, reflecting the fact that prior uncertainty about the exclusion restriction increases the dispersion of the posterior distribution of β. This increase in posterior uncertainty regarding β is of course higher the greater is prior uncertainty regarding the exclusion restriction. Consider for example when all three sample correlations are equal to 0.5 and the sample size is equal to 100. When $\eta=10$, corresponding to significant uncertainty about the exclusion restriction, the 95 percent confidence interval for β is 2.14 times larger than the

benchmark case where $\rho=\phi=0$ by assumption. However, as η increases this magnification of posterior uncertainty is smaller, and when $\eta=500$ the confidence intervals are just 1.03 times larger than the benchmark case.

Unsurprisingly, Table 2 also confirms that in all cases the magnification of posterior uncertainty is greater the larger is the sample size. For example, when all three sample correlations are equal to 0.5 and the sample size is equal to 100, the confidence interval for β is inflated by a factor of 2.14 when T=100, but it is inflated by a factor of 4.45 when T=500. The reason for this is the same as in Section 2 in the OLS case. There we saw that the correction to the posterior variance of β to capture uncertainty about the exclusion restriction does not decline with sample size, and so its effect on posterior uncertainty is proportionately greater the larger is the sample size.

The more interesting insight from Table 2 is that the magnification of posterior uncertainty about β also depends on the moments of the observed sample in a very intuitive way. Consider the first panel of Table 2, where I vary the strength of the first-stage sample correlation between the instrument and the endogenous variable, R_{xz}, holding constant the other two correlations.[8] In the standard case where $\rho=\phi=0$ by assumption, the confidence intervals of course shrink as the strength of the first-stage relationship increases. However, the magnification of the posterior variance *increases* as the strength of the first-stage relationship increases. The intuition for this is analogous to the intuition for the effects of sample size. A larger sample size, and also a stronger first-stage relationship between the instrument and the endogenous variable permit more precise inferences about β. However, a larger sample size and a stronger first-stage regression cannot reduce our intrinsic uncertainty about the validity of the exclusion restriction, and so the adjustment to the posterior variance to account for this is proportionately greater. Of course this does not mean that uncertainty about the

[8] In these examples I have chosen hypothetical samples in which we are unlikely to encounter well-known weak-instrument pathologies. In fact, the minimum correlation of 0.3 between the endogenous variable and the instrument in this table is deliberately chosen to ensure that the first-stage F-statistic is almost 10 in the smallest sample of size T=100 that I consider, and is greater than 10 in all other cases. This corresponds to the rule of thumb proposed by Staiger and Stock (1997) for distinguishing between weak and strong instruments. These weak-instrument pathologies pose no particular difficulties for Bayesian analysis that bases inference on the entire posterior distribution of β. However, with weak instruments the Bayesian highest posterior density intervals I focus would no longer necessarily be symmetric around the mode of the posterior distribution.

exclusion restriction is less important in an absolute sense in small samples or with weak instruments -- only that its effects on posterior uncertainty are smaller *relative to* other sources of posterior imprecision about the parameters of interest.

The same insight holds in the second and third panels of Table 2. In the second panel, I vary the strength of the observed sample correlation between the dependent variable and the instrument, R_{yz}. Since I am holding constant the other two correlations in this panel, larger values of R_{yz} correspond to greater endogeneity problems, and hence less precise IV estimates of the structural slope coefficient β in the benchmark case where $\rho=\phi=0$ by assumption. Since varying the extent of the endogeneity problem does not affect the intrinsic uncertainty about the exclusion restriction, I find that the magnification of the confidence interval declines as R_{yz} increases. A similar effect occurs in the third panel, where I vary the strength of the observed correlation between y and x. Since I am holding the other two correlations constant, higher values of R_{xy} imply a more precisely-estimated structural relationship between these two variables. However, once again this does not affect intrinsic prior (and posterior) uncertainty about the exclusion restriction, and so the magnification of the confidence intervals increases as R_{xy} increases.

In summary, we have seen that prior uncertainty about the exclusion restriction can substantially increase posterior uncertainty about the key structural slope coefficient of interest, β. The magnitude of this inflation of posterior uncertainty depends of course depends on the degree of prior uncertainty about the exclusion restriction. But it also depends on the characteristics of the observed sample in a very intuitive way. Holding other things constant, a greater sample size, a stronger first-stage relationship between the instrument and the endogenous variable, a stronger structural correlation between dependent variable and the endogenous variable, and a weaker reduced-form correlation between the endogenous variable and the instrument all imply a more precise IV estimator, absent any prior uncertainty about the exclusion restriction. However, since none of these factors help to reduce prior (or posterior) uncertainty about the exclusion restriction, this uncertainty becomes relatively more important.

4. Empirical Applications

I next demonstrate the quantitative importance for inference of prior uncertainty about exclusion restrictions in three well-known empirical studies that use linear instrumental variables models. Acemoglu, Johnson and Robinson (2001, hereafter AJR) study the causal effects of institutions on economic development. Using a sample of 64 former colonies, they regress the logarithm of GDP per capita on a measure of property rights protection. They propose using historical data on mortality rates experienced by settlers during the colonial period as a novel instrument for institutional quality. AJR argue that in areas where settlers experienced high mortality rates, colonial powers had few incentives to set up institutions that protect property rights and provide a foundation for subsequent economic activity. In a simple bivariate specification there are a number of obvious concerns regarding the validity of the exclusion restriction that settler mortality rates matter for development only through their effects on institutional quality. Historical settler mortality rates might be correlated with the tropical location and intrinsic disease burden of a country, and these factors may matter directly for modern development. AJR seek to address such concerns in their paper through the addition of various control variables to capture these effects. For example, we will show results using one of their core specifications in which they control for latitude to capture such locational effects (Table 4, Column 2 in AJR). And in the paper they also present a wide range results with direct controls for location and the disease burden.[9]

Nevertheless, a reader of AJR might reasonably entertain some doubts as to whether the exclusion restriction holds exactly even in these extended specifications. There are many potential correlates of settler mortality rates that might in turn be correlated with development outcomes. For example, Glaeser et. al. (2004) argue that low settler mortality rates may have operated through investments in human capital rather than institutions to protect property rights. Here we do not take any stand as to

[9] Ideally I would like to use one of AJR's specifications with a more complete set of control variables to illustrate the effects of uncertainty about exclusion restrictions. However, in many of their specifications with more control variables, their instruments are much weaker, and I do not want to conflate my point about uncertainty regarding exclusion restrictions with the well-known concerns with weak instruments. For example, in Columns (7) and (8) of Table 4, AJR introduce continent dummies, and continent dummies together with latitude. In these specifications, I find first-stage F-statistics on the excluded instrument of 6.83 and 3.97, well below the Staiger and Stock (1997) rule of thumb of 10. This suggests that the settler mortality instrument does not have sufficiently strong explanatory power within geographic regions.

which of these potential failures of the exclusion restriction is the right one. Rather we simply argue that reasonable people might question whether the exclusion restriction holds exactly, and might entertain some probability that it is not in fact true.

My second example is Frankel and Romer (1999, hereafter FR), who study the relationship between trade openness and development in a large cross-section of countries. They regress log GDP per capita on trade as a share of GDP. To address concerns about potential reverse causation and omitted variables, they propose a novel instrument based on the geographical determinants of bilateral trade. In particular, they estimate a regression of bilateral trade between country pairs on the distance between the countries in the pair, their size measured by log population and log area, and a dummy variable indicating whether either country in the pair is landlocked. They then use the fitted values from this bilateral trade regression to come up with a constructed trade share for each country that reflects only these geographical determinants of trade. They then use this as an instrument for trade. In their core specification, they also control directly for country size, as measured by log population and log land area, to control for the problem that large countries tend to trade less and these size variables also enter in the bilateral trade equation. There are however various reasons why the necessary exclusion restriction (that the geographically-determined component of trade matters for development only through its effects on overall trade) may not hold exactly. For example, Rodriguéz and Rodrik (2000) discuss various channels through which the geographical variables in the FR bilateral trade regression might have direct effects on per capita incomes.

My third example comes from Rajan and Zingales (1998, hereafter RZ), who study the relationship between financial development and growth. In contrast with the previous two papers that exploit purely cross-country variation, this paper uses a novel identification strategy that exploits within-country cross-industry differences in manufacturing growth rates. They construct a measure of the dependence of different manufacturing sectors on financial services, and then ask whether industries that are more financially-dependent grow faster in countries where financial development is greater. In particular, they estimate regressions of the growth rate of industry i in country j on a set of country dummies, a set of industry dummies, the initial size of the industry, and an interaction of the financial dependence of the sector with the level of financial

development in the country. In a number of specifications, RZ instrument for this final interaction term with variables capturing the legal origins of the country and a measure of institutional quality, all interacted with a measure of financial development. In particular, I will focus on the specification in Table 4, column 6 of RZ, where the relevant measure of financial dependence is an index of accounting standards recording the types of information provided in annual reports of publicly-traded corporations in a cross-section of countries.

This third example differs from the previous ones in two key respects. First, because RZ rely on the within-country variation in sectoral growth rates, potential violations of the exclusion restriction are less obvious than in the previous two cases. In RZ, the requirement is that the instruments be orthogonal to the country- and industry-specific component of growth, since the regressions contain country and industry dummies. Thus for example, concerns about the exclusion restriction are not that countries with faster growth adopt better accounting standards, but rather that countries with a relatively faster growth in financially-dependent industries would adopt better accounting standards. Nevertheless there might be residual concerns about the validity of the exclusion restriction in this case. The second difference is that RZ use multiple instruments, while the results I show above apply to the case of a single instrument. To make the RZ results fit into the framework of this paper, I choose just one of their instruments and first reproduce the RZ results in this just-identified case. For this purpose I choose their index of efficiency and integrity of the legal system, produced by a commercial risk rating agency, as the one instrument of choice. Doing so gives a result that is of comparable significance to the RZ core result, although the magnitude of the estimated coefficient becomes somewhat larger than what RZ report.[10]

I use datasets provided by the authors to reproduce their results. In each of the three examples, I first project the dependent variable, the regressor of interest, and the instrument on all the remaining control variables that these authors treat as exogenous, so that I can identify these residuals as y, x, and z in the discussion above. I also normalize the variance of z to be equal to one, consistent with the discussion above. I

[10] An alternative is to use just their dummy variable for Scandinavian legal origins as an instrument, which generates results that are quite similar to those reported by RZ. Conversely, using either dummies for British or French legal origins alone as an instrument does not deliver significant IV estimates of the coefficient on the interaction variable of interest.

then take 10,000 draws from the posterior distribution of β, for alternative values of η corresponding to varying degrees of prior uncertainty about the exclusion restriction. I then compute the 2.5th, 50th and 97.5th percentiles of this distribution.

Table 3 summarizes the results, with three panels corresponding to the three examples. In each panel in the first column I report the sample size and my replication of the relevant IV slope coefficient and standard error from each paper. In the columns of the table I provide summary statistics on the posterior distribution for the slope coefficient, for varying degrees of prior uncertainty about the exclusion restriction. In addition, Figure 2 plots the posterior densities for the slope coefficient for selected values of η. Unsurprisingly, in all three panels of this figure we clearly see how the posterior distribution of the slope coefficient becomes more dispersed as uncertainty about the exclusion restriction increases.

This increase in posterior dispersion is quantified in the table, which reports the 2.5th, 50th, and 97.5th percentiles of the posterior distribution of the structural slope coefficient for each of the three papers. To read this table, it is useful to begin with the last column which reports these percentiles for the limiting case where η tends to infinity and thus the prior distribution imposes ϕ=0 with certainty. This corresponds to the standard Bayesian IV estimates in which there is no uncertainty regarding the exclusion restriction. Because of my choice of diffuse priors for all of the parameters other than ϕ, when ϕ=0 these Bayesian results mimic the classical ones quite closely, with these percentiles quite similar to the 95 percent confidence intervals reported in the first column. This is particularly so for RZ, while for FR and AJR the posterior distribution of the slope is somewhat longer right tail, with the result that the 97.5th percentiles are a bit higher than the upper bounds of the classical confidence intervals. This is also apparent in Figure 2, where the thin solid line plots a normal distribution with mean and standard deviation corresponding to the classical IV slope coefficient estimate and estimated standard error. For RZ this normal distribution coincides almost perfectly with the posterior distribution for the slope when ϕ=0, while there are some small discrepancies for the other two papers.

Moving from right to left in Table 3 illustrates the effects of greater prior uncertainty about the exclusion restriction. In each of the three panels, I summarize this

increase in the dispersion of the posterior distribution by reporting the length of the interval from the 2.5th percentile to the 97.5th percentile, relative to the length of the same interval when ϕ=0 with certainty. These intervals expand substantially as uncertainty about the exclusion restriction increases. For example, for FR in the middle panel, this interval is 2.8 times as wide when η=10, while for RZ in the bottom panel it is 7.26 times as wide. This greater proportional effect on posterior uncertainty about the structural slope is consistent with what we saw in the artificial samples in Table 2, as RZ have a larger sample size and a stronger instrument than do FR. In contrast, for AJR with their smaller sample, the increase in posterior dispersion is smaller.

Table 2 also can be used to determine how great prior uncertainty about the exclusion restriction needs to be in order for the interval from the 2.5th percentile to the 97.5th percentile of the posterior distribution of β to include zero. In the case of AJR, their particular specification that we report is most robust to uncertainty about the exclusion restriction. Even when η=5, so that there is a great deal of prior uncertainty, with 90 percent of the prior probability mass for ϕ (and ρ) between -0.46 and 0.46, the 2.5th percentile of the posterior distribution of the slope is greater than zero. This is not however the case for FR and RZ. Moving from η=200 to η=100, the 2.5th percentile of the posterior distribution of the slope falls below zero. This in turn means that if the prior distribution of ϕ (and ρ) is such that more than 10 percent of the prior probability mass falls outside the interval of about (-0.1,0.1), then the Bayesian analog of the 95 percent confidence interval includes zero.

5. Extensions and Conclusions

The validity of the IV estimator depends crucially on the validity of fundamentally untestable exclusion restrictions. Typically these exclusion restrictions are assumed to hold exactly in the relevant population. However, in many empirical examples it is reasonable to doubt their validity. In this paper I have shown how to explicitly incorporate prior uncertainty about the exclusion restriction into the linear IV regression model. This prior uncertainty about the exclusion restriction leads to greater posterior uncertainty about parameters of interest, in some cases quite substantially so. This enables straightforward checks of the robustness of inferences about structural parameters to varying degrees of prior uncertainty about the exclusion restriction.

There are at least two natural extensions of the results presented here. The first I have already discussed: allowing the prior distribution for the correlation between the instrument and the error term to have a non-zero mean. This would encompass not only prior uncertainty about the validity of the exclusion restriction, but also prior beliefs about the direction of likely violations of the exclusion restriction. For example, one might specify a prior distribution for ϕ that is a translation of a beta distribution, i.e. $(\phi+1)/2 \sim \text{Beta}(\eta_1,\eta_1)$. With appropriate choices of the prior parameters η_1 and η_2, a prior such as this can capture prior beliefs regarding both the mean and the variance of ϕ. Since there is no updating of the prior distribution of ϕ, we will have the same posterior distribution, and we can simply (numerically) integrate over this distribution to arrive at the marginal posterior distribution for the slope coefficients of interest. This will have predictable effects on the results presented here: the posterior mean of the distribution of the structural slope coefficients will need to be adjusted to reflect the non-zero prior and posterior mean for the distribution of ϕ, since the expectation in Equation (19) will no longer be zero. While this extension may be practically useful in many situations where there might be obvious potential directions for violations of the exclusion restriction, conceptually this adds little in the way of additional insights.

The second is to consider the case of multiple instruments and multiple endogenous variables. In this paper, I have focused on the case of a single endogenous variable and a single instrument in order to keep the results as transparent as possible. Moving to the case of multiple endogenous variables and potential overidentification also

poses no particular conceptual problems, although it does pose two modest practical difficulties. First, when there are multiple instruments, we need to elicit a prior distribution over the correlation between each of the instruments and the structural error term, rather than just a simple univariate prior over a single parameter that I have used here. In practice, it may be difficult to flexibly specify such a prior in a way that captures differing degrees of certainty about the exclusion restriction for each instrument. Second, in the case of overidentification, the mapping from the reduced-form parameters to the structural parameters is more complex, and therefore it is more difficult to simulate the prior and posterior distribution of the structural parameters that is implied by the prior and posterior distribution over the reduced-form parameters. Hoogerheide, Kleibergen and Van Dijk (2007) provide further details on this case.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson (2001). "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*. 91(5):1369-1401.

Berkowtiz, Daniel, Mehmet Caner and Ying Fang (2008). "Are 'Nearly Exogenous' Instruments Reliable?". Economics Letters. (article in press).

Conley, Tim, Christian Hansen and Peter E. Rossi (2007). "Plausibly Exogenous". Manuscript. Graduate School of Business, University of Chicago.

Frankel, Jeffrey A. and David Romer (1999). "Does Trade Cause Growth?" *The American Economic Review*, (June) 379-399.

Glaeser, Edward, Rafael Laporta, Florencio Lopez-de-Silanes, and Andrei Shleifer (2004). "Do Institutions Cause Growth?". Journal of Economic Growth. 9(3):271-303.

Hahn, Jinyong and Jerry Hausman (2006). "IV Estimation with Valid and Invalid Instruments". Annales d'Economie et Statistique.

Hoogerheide, Lennart, Frank Kleibergen and Herman van Dijk (2007). "Natural Conjugate Priors for the Instrumental Variables Regression Model Applied to the Angrist-Krueger Data". 138(1):63-103.

Murray, Michael (2006). "Avoiding Invalid Instruments and Coping with Weak Instruments". Journal of Economic Perspectives. 20(4):111-132.

Poirier, Dale J. (1998). "Revising Beliefs in Nonidentified Models". Econometric Theory. 14:483-509.

Rajan, Raghuram and Luigi Zingales (1998). "Financial Dependence and Growth". American Economic Review. 88(3):559-586.

Rodriguez, Francisco and Dani Rodrik (2001). "Trade Policy and Economic Growth: A Skeptic's Guide to the Cross-Country Evidence". NBER Macroeconomics Annual. 15:261-325.

Small, Dylan (2007). "Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions". Journal of the American Statistical Association. 102(479):1049-1058.

Staiger, D. and J.H. Stock (1997). "Instrumental Variables Regression With Weak Instruments". Econometrica. 65:557-586.

Table 1: Inference in the OLS Case

	Value of Prior Parameter η					
	5	10	100	200	500	1000
90% Prior Probability of ρ Between:						
Lower	-0.46	-0.34	-0.12	-0.08	-0.05	-0.04
Upper	0.46	0.34	0.12	0.08	0.05	0.04
Inflation of Posterior Standard Deviation of β When T=:						
100	3.32	2.45	1.22	1.12	1.05	1.02
200	4.58	3.32	1.41	1.22	1.10	1.05
500	7.14	5.10	1.87	1.50	1.22	1.12
1000	10.05	7.14	2.45	1.87	1.41	1.22

Table 2: Inference in the IV Case

				Value of Prior Parameter η				
				5	10	100	200	500
90 percent of prior probability between:								
			Lower	-0.46	-0.34	-0.12	-0.08	-0.05
			Upper	0.46	0.34	0.12	0.08	0.05
Assumptions on Observed Sample				**Width of 95% Confidence Interval for Indicated Value of η** (Relative to Width when ϕ=0)				
Vary Strength of First-Stage CORR(x,z)								
Rxy= 0.5	*Ryz= 0.5*	***Rxz= 0.3***						
95% CI for β	(1.00, 4.02)		*T=100*	1.85	1.49	1.05	1.05	1.04
95% CI for β	(1.32, 2.20)		*T=500*	4.44	3.16	1.41	1.26	1.11
Rxy= 0.5	*Ryz= 0.5*	***Rxz= 0.5***						
95% CI for β	(0.65, 1.52)		*T=100*	2.95	2.14	1.17	1.10	1.03
95% CI for β	(0.84, 1.19)		*T=500*	6.42	4.45	1.72	1.42	1.17
Rxy= 0.5	*Ryz= 0.5*	***Rxz= 0.7***						
95% CI for β	(0.47, 0.99)		*T=100*	3.28	2.41	1.22	1.13	1.05
95% CI for β	(0.61, 0.83)		*T=500*	7.19	5.00	1.86	1.48	1.23
Vary Strength of Reduced Form CORR(y,z)								
Rxy= 0.5	***Ryz= 0.3***	*Rxz= 0.5*						
95% CI for β	(0.24, 0.99)		*T=100*	3.71	2.64	1.24	1.11	1.05
95% CI for β	(0.45, 0.76)		*T=500*	8.09	5.75	2.01	1.60	1.29
Rxy= 0.5	***Ryz= 0.5***	*Rxz= 0.5*						
95% CI for β	(0.65, 1.52)		*T=100*	2.95	2.14	1.17	1.10	1.03
95% CI for β	(0.84, 1.19)		*T=500*	6.42	4.45	1.72	1.42	1.17
Rxy= 0.5	***Ryz= 0.7***	*Rxz= 0.5*						
95% CI for β	(1.01, 2.11)		*T=100*	2.02	1.57	1.07	1.06	1.05
95% CI for β	(1.21, 1.65)		*T=500*	4.17	3.13	1.33	1.21	1.09
Vary Strength of Structural CORR(y,x)								
Rxy= 0.3	*Ryz= 0.5*	*Rxz= 0.5*						
95% CI for β	(0.60, 1.63)		*T=100*	2.57	1.92	1.12	1.06	1.02
95% CI for β	(0.81, 1.23)		*T=500*	5.34	3.70	1.53	1.31	1.13
Rxy= 0.5	*Ryz= 0.5*	*Rxz= 0.5*						
95% CI for β	(0.65, 1.52)		*T=100*	2.95	2.14	1.17	1.10	1.03
95% CI for β	(0.84, 1.19)		*T=500*	6.42	4.45	1.72	1.42	1.17
Rxy= 0.7	*Ryz= 0.5*	*Rxz= 0.5*						
95% CI for β	(0.72, 1.39)		*T=100*	3.65	2.70	1.28	1.16	1.07
95% CI for β	(0.87, 1.15)		*T=500*	8.14	5.73	2.03	1.64	1.28

Table 3: Empirical Examples

	Value of Prior Parameter η					
	5	10	100	200	500	∞
90% Prior Probability of ρ Between:						
Lower	-0.46	-0.34	-0.12	-0.08	-0.05	0.00
Upper	0.46	0.34	0.12	0.08	0.05	0.00
Acemoglu-Johnson-Robinson (2001)						
(Table 4, Column 2)						
T=64						
IV Slope = 0.96						
IV Standard Error = 0.21						
95% C.I. = (0.53, 1.39)						
Posterior Distribution for Slope						
2.5th Percentile	0.08	0.32	0.61	0.63	0.63	0.65
Mode	0.95	0.96	0.96	0.96	0.96	0.96
97.5th Percentile	2.31	2.06	1.80	1.81	1.76	1.75
Increase in P025-P975 range	2.02	1.57	1.08	1.07	1.02	1.00
Frankel-Romer (1999)						
(Table 3, Column 2)						
T=150						
IV Slope = 1.97						
IV Standard Error = 0.91						
95% C.I. = (0.18, 3.76)						
Posterior Distribution for Slope						
2.5th Percentile	-5.62	-3.61	-0.28	0.01	0.14	0.31
Mode	1.98	1.95	1.97	1.98	1.96	1.96
97.5th Percentile	10.73	8.66	5.37	5.04	4.83	4.69
Increase in P025-P975 range	3.73	2.80	1.29	1.15	1.07	1.00
Rajan-Zingales (1998)						
(Table 4, Column 6)						
T=1067						
IV Slope = 0.31						
IV Standard Error = 0.08						
95% C.I. = (0.16, 0.46)						
Posterior Distribution for Slope						
2.5th Percentile	-1.27	-0.82	-0.06	0.02	0.10	0.16
Mode	0.30	0.31	0.31	0.31	0.31	0.31
97.5th Percentile	1.85	1.41	0.70	0.60	0.53	0.47
Increase in P025-P975 range	10.14	7.26	2.45	1.90	1.40	1.00

Figure 1: The Prior Distribution for ρ, OLS Case

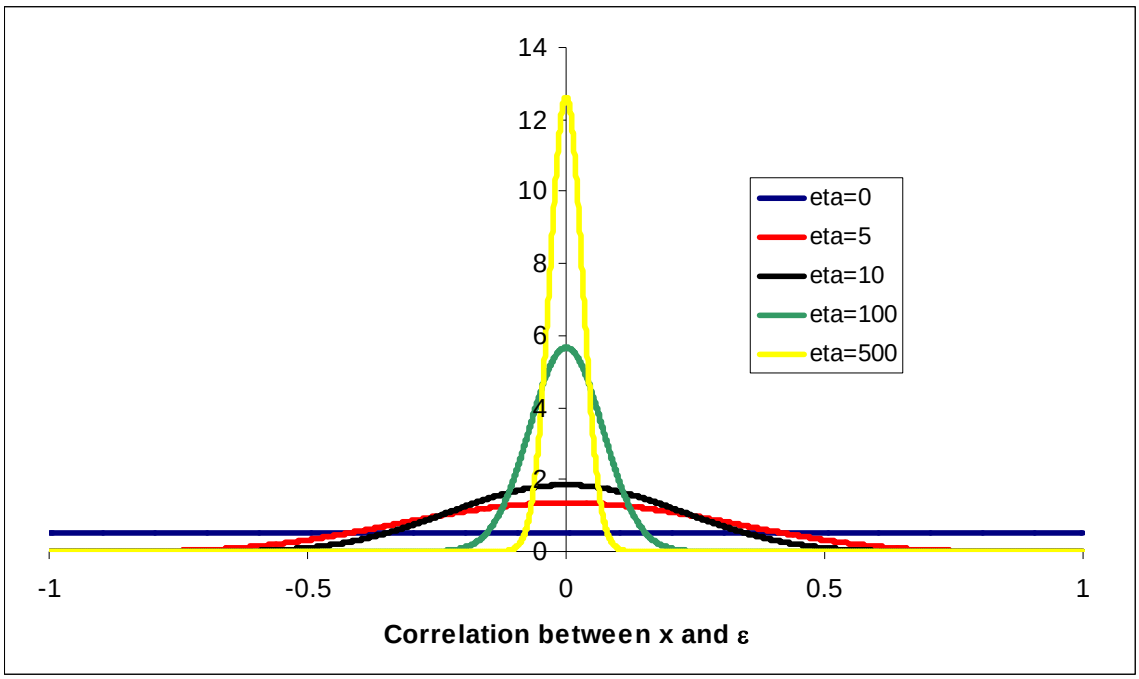

Figure 2: Posterior Distribution for Structural Slopes